SENTI-MD: IMDb Movie Review Sentiment Analysis Summary

Team/Student: [Prateek P-PES1UG23AM211, Vivian Philip-PES1UG23AM907]

1. Problem & Goal

The core objective was to build a highly accurate **Machine Learning (ML) classifier** to automatically determine the sentiment of IMDb movie reviews. The challenge lies in classifying unstructured text (reviews) into a binary output: **Positive (1) or Negative (0)**, overcoming the ambiguity and complexity inherent in human language.

2. Methodology and Implementation

A. Data Preparation

- **Dataset:** IMDb Movie Review Dataset (50,000 reviews).
- **Preprocessing:** Cleaning (removing HTML, punctuation), lowercasing, and removal of common **stop words** (NLTK).
- **Feature Extraction:** Text was vectorized using **TF-IDF (Term Frequency-Inverse Document Frequency)** with a focus on **3-grams** to capture local context and phrases (e.g., "not good").

B. Model Selection

We evaluated five different classifiers (Logistic Regression, SVM, Linear SVC, etc.). **Logistic Regression** consistently outperformed the others on the high-dimensional TF-IDF feature set, offering the best balance of performance and efficiency.

C. Implementation

The project was entirely implemented in **Python** using the **scikit-learn** and **NLTK** libraries within a **Google Colab** environment.

3. Key Results & Conclusion

The final evaluation was performed on a held-out test set, confirming the model's strong generalization ability.

Metric	Model	Result
Best Classifier	**Logistic Regression**	**0.884**
Feature Method	**TF-IDF (3-grams)**	**~0.88**
Accuracy	**0.884**	**N/A**
F1-Score	**N/A**	**~0.88**

Conclusion: The **SENTI-MD** pipeline, anchored by the **Logistic Regression** model and **TF-IDF (3-gram)** features, achieved a high classification accuracy of **88.4%**. This result validates the chosen NLP approach as highly effective for large-scale, automated sentiment classification of movie reviews.